Introduction to R

Noushin Ghaffari, PhD
Bioinformatics Scientist, Genomics and Bioinformatics, Texas A&M AgriLife Research
Research Scientist, Texas A&M High Performance Research Computing
What is R?

- Open source programming language and software environment for statistical computing and graphics
- Supported by the R Foundation for Statistical Computing
- Supports multiple platforms and can be easily extended.

- The “Comprehensive R Archive Network” (CRAN) is a collection of sites which carry identical material, consisting of the R distribution(s), the contributed extensions, documentation for R, and binaries.
- The CRAN master site at WU (Wirtschaftsuniversität Wien) in Austria can be found at the URL https://CRAN.R-project.org/
  - Mirrored daily to many sites around the world https://CRAN.R-project.org/mirrors.html
Why R?

- Open source
- Comprehensive collection of statistical functions
  - Linear modeling, classification, clustering, genomics analysis, economic and financial analysis, etc.
- Large user community – support
- Collection of “packages”
  - A package is a shared code, documentation/Vignettes, data (occasionally)
- Multi platform, but not too sensitive about the source platform
- Command-line or graphical user interface (recently)
- Anyone can contribute
What is RStudio?

- An integrated development environment (IDE) for R
- Graphical Interface – user friendly environment
- Embedded Text Editor
- Auto completion on names (functions)
- Simplified plot and output view
Running R for this Course

• Your own laptop - You need to install R on your laptop using R download page: https://cran.r-project.org/mirrors.html

• TAMU Open Access Lab (OAL) computers have R and RStudio installed

• TAMU NetID - You can use the online Jupyter notebook for the class
  • Off campus needs VPN

• HPRC Portal to run Rstudio: http://portal.hprc.tamu.edu

• TAMU HPRC Ada System Account – You can use R_Tamu on Ada or choose Jupyter Notebook
  • Off campus needs VPN
Using Jupyter Notebook

**Jupyter** is a web application that allows you to easily connect to a remote server and run applications. We will be using **Jupyter** to connect to **titan.tamu.edu** and use R.

To access the **Jupyter Notebook** you will need:

- An Internet-connected device (desktop, laptop, or large tablet is preferred)
- Internet access (tamulink-wpa Wi-Fi, OAL Ethernet)
- An Internet browser
Jupyter Notebook Access

(1) Open your Internet browser.

(2) Go to one of the following links: https://titan.tamu.edu:8000/

(3) You should then see a login window. Use your NetID credentials to log in. This is the same username and password as Howdy and Ada.
(4) Upon successful login you will see some files and directories. Click on the “"Introduction_to_R_HPRC_TAMU_April2018.ipynb”

(5) Before class begins, please restart your kernel and clear all outputs. To do this, click on “Kernel -> Restart & Clear Output”
Connecting to HPRC to Use R

- **SSH (secure shell)**
  - The only program allowed for remote access; encrypted communication; freely available for Linux/Unix and Mac OS X hosts;
- **For Microsoft Windows PCs, use MobaXterm**
  - https://hprc.tamu.edu/wiki/HPRC:MobaXterm
    - You are able to view images and use GUI applications with MobaXterm
  - or **Putty**
    - https://hprc.tamu.edu/wiki/HPRC:Access#Using_PuTTY
      - You can not view images or use GUI applications with PuTTY
- Both state of Texas law and TAMU regulations prohibit the sharing and/or illegal use of computer passwords and accounts
- Don’t write down passwords
- Don’t choose easy to guess/crack passwords
- Change passwords frequently
Using SSH - MobaXterm (on Windows)

- Important policy information
- Unauthorized use of HPRC resources is prohibited and subject to criminal prosecution.
- Use of HPRC resources in violation of United States export control laws and regulations is prohibited. Current HPRC staff members are US citizens and legal residents.
- Sharing HPRC account and password information is in violation of State Law. Any shared accounts will be disabled.
- Authorized users must also adhere to all policies at:
  - Website: [http://hprc.tamu.edu](http://hprc.tamu.edu)

**Message of the Day**

- WARNING: There are NO active backups of user data.!!
- Please restrict usage to 8 CORES across ALL Ada login nodes. Users found in violation of this policy will be SUSPENDED.

**Ada Scheduled Maintenance Completed**

The maintenance for Ada has been completed. Batch job scheduling has resumed.

Your current disk quotas are:

- **Disk**
  - Limit:
    - File Usage: 117.2 MB of 100 GB
    - Disk Usage: 3419 MB of 10000 GB
- **scratch**
  - Limit:
    - File Usage: 6.8 MB of 200000 GB
    - Disk Usage: 1 MB of 100 GB
- **tiered**
  - Limit:
    - File Usage: 0 MB of 50000 GB
    - Disk Usage: 0 MB of 100 GB

Type 'showquota' to view these quotas again.

`whomps@login5:~$`
Using SSH to Access Ada

```bash
ssh -X user_NetID@ada.tamu.edu
```

https://hprc.tamu.edu/wiki/Ada:Access

You may see something like the following the first time you connect to the remote machine from your local machine:

Host key not found from the list of known hosts.
Are you sure you want to continue connecting (yes/no)?

Type yes, hit enter and you will then see the following:

Host 'ada.tamu.edu' added to the list of known hosts.
user_NetID@ada.tamu.edu's password:
Login and Set up

- Login to Ada using SSH or MobaXterm
- Let’s take a look at the path and create appropriate directories

```
echo $SCRATCH
cd $SCRATCH
Pwd
mkdir Intro_to_R_Apr18
```
Transferring the Code

The R code and its submission script should be copied to users’ local directory

```
cd $SCRATCH/Intro_to_R_Apr18

cp /scratch/training/Intro_to_R/Apr18/* .
```
#randomly generates 10000 number in a matix[100,100]
rand_data <- matrix(ncol=100,nrow=100,data=runif(10000,1,100))

#getting log2 of the data and transposing the results
rand_data_t_log2 <- t(log2(rand_data))

#calculating the min, max and average for rows and columns
mean_cols <- apply(rand_data,2,mean)
mean_rows <- apply(rand_data,1,mean)
min_rows <- apply(rand_data,1,min)
min_cols <- apply(rand_data,2,min)
max_rows <- apply(rand_data,1,max)
max_cols <- apply(rand_data,2,max)
#making plot and saving results in a pdf file

pdf("Rand_Data_QC.pdf")
g_range <- range(0,min_rows,max_rows)
plot(mean_rows, type="l", col="blue", ylim=g_range, ann=FALSE)
lines(mean_cols, type="l", col="red")
lines(max_rows, type="l", col="green")
lines(max_cols, type="l", col="purple")
lines(min_rows, type="l", col="royalblue")
lines(min_cols, type="l", col="coral")
legend(80,40,c("Mean Rows","Mean Cols","Max Rows","Max Cols","Min Rows","Min Cols"),col=c("blue","red","green","purple","coral","royalblue"),lty=1,cex=0.55)
dev.off()
# saving data files
write.csv(rand_data,"rand_data.csv")
write.csv(rand_data_t_log2,"rand_data_t_log2.csv")
R_Script_Submit.sh

#BSUB -J Testing_R_Script  # sets the job name to Testing_R_Script.
#BSUB -L /bin/bash        # uses the bash login shell to initialize the job's execution environment.
#BSUB -W 1:00             # sets to 5 hours the job's runtime wall-clock limit.
#BSUB -n 1                # assigns 1 core for execution.
#BSUB -R "span[ptile=1]"  # assigns 1 core per node.
#BSUB -R "rusage[mem=5000]" # reserves ~5GB per process/CPU for the job
#BSUB -M 5000             # sets to ~5GB the per process enforceable memory limit.
#BSUB -o stdout.%J        # directs the job's standard output to stdout.jobid

## Load the necessary modules
module purge
module load R_tamu/3.4.2-iomkl-2017A-Python-2.7.12-default-mt

## Launch R with proper parameters
Rscript Data_Generator.R
Submitting the RScript to Ada

```bash
bsub < $SCRATCH/Intro_to_R_Apr18/R_Script_Submit.sh
Bjobs
ls -l
```
Course Content

Parts of this course are based on Software Carpentry "Programming with R" and "R for Reproducible Scientific Analysis" lessons.

“Software Carpentry is a volunteer non-profit organization dedicated to teaching basic computing skills to researchers.”

- https://software-carpentry.org/lessons/
- Reference page for R lessons (cheat sheet)
Any question?
nghaffari@tamu.edu